Elsevier Editorial System(tm) for Gene

Manuscript Draft

Manuscript Number:

Title: The properties of CpG islands in the putative promoter regions of human immunoglobulin (Ig) genes

Article Type: Research Paper

Section/Category: Functional Genomics

Keywords: CpG island, Human genome, Mouse genome, CpG density, CpG distribution pattern

Corresponding Author: Dr Guang Bin Liu, PhD

Corresponding Author's Institution: University of Queensland

First Author: Guang B Liu, PhD

Order of Authors: Guang B Liu, PhD; Rong Chen, Master; Ya F Jiang, PhD; Hong Yan, PhD; John D Pettigrew, PhD; Kong-Nan Zhao, PhD

Manuscript Region of Origin:

Abstract: CpG is a GC rich motifs in the gene promoter region, which can play important roles in gene silencing and imprinting. This study examines the properties of CpG islands in the putative promoter regions (PPRs) of human and mouse immunoglobulin (Ig) genes. The PPRs of both human and mouse Ig genes irrespective of gene chromosomal localization are apparently CpG island poor, with low percentage of the CpG islands overlapped with the transcription start site (TSS). The human Ig genes that have CpG islands in the PPRs show a very narrower range of CpG densities. 47 % of these Ig genes fall in the density range of 3.5-4 CpG/100bp, only 4.2% have the CpG islands with density ranging 6.1-8 CpG/100bp. CpG distributions within CpG islands can be classified into five patterns: Pat A, B, C, D, and E. 21.6% and 10.8% of the Ig genes show their CpG distributions in Pat B and Pat D; but only 8.2% and 3.8% of the non-Ig genes

have the CpG distributions in the two patterns. Moreover, the length of CpG islands is shorter in human Ig genes than in non-Ig genes, but much longer than in mouse orthologous. Thus, our data suggest that the occurrence and distribution of CpG islands in PPRs of human and mouse Ig genes is nonneutral and nonrandom, which may reflect the species and gene specification.

Abstract:

CpG is a GC rich motifs in the gene promoter region, which can play important roles in gene silencing and imprinting. This study examines the properties of CpG islands in the putative promoter regions (PPRs) of human and mouse immunoglobulin (Ig) genes. The PPRs of both human and mouse Ig genes irrespective of gene chromosomal localization are apparently CpG island poor, with low percentage of the CpG islands overlapped with the transcription start site (TSS). The human Ig genes that have CpG islands in the PPRs show a very narrower range of CpG densities. 47 % of these Ig genes fall in the density range of 3.5-4 CpG/100bp, only 4.2% have the CpG islands with density ranging 6.1-8 CpG/100bp. CpG distributions within CpG islands can be classified into five patterns: Pat A, B, C, D, and E. 21.6% and 10.8% of the Ig genes show their CpG distributions in Pat B and Pat D; but only 8.2% and 3.8% of the non-Ig genes have the CpG distributions in the two patterns. Moreover, the length of CpG islands is shorter in human Ig genes than in non-Ig genes, but much longer than in mouse orthologous. Thus, our data suggest that the occurrence and distribution of CpG islands in PPRs of human and mouse Ig genes is nonneutral and nonrandom, which may reflect the species and gene specification.

The properties of CpG islands in the putative promoter regions of human immunoglobulin (Ig) genes

Key words: Human genome, Mouse genome, CpG density, CpG distribution pattern

Guang B. Liu[a]*, Rong Chen[b], Ya F. Jiang[c], Hong Yan[b], John D. Pettigrew[a] and Kong-Nan Zhao[d]

Author affiliation:
a Vision, Touch and Hearing Research Centre, Faculty of Biomedical Sciences, The University of Queensland, Brisbane, Australia
b Department of Health Statistics, School of Medicine, Xi'an Jiaotong University, Xi'an, P.R. China
c School of Nursing, Peking Union Medical College, Ba Da Chu Road, Beijing, P. R. China
d Centre for Immunology and Cancer Research, The University of Queensland, Princess Alexandra Hospital, Brisbane, Australia.

* Corresponding author
Corresponding author: Guang B. Liu
Vision Touch and Hearing Research Centre, Faculty of Biomedical Sciences
The University of Queensland, Brisbane, Australia
Phone number: 617 3365 4072
Facsimile: 617 3365 4522
e-mail: g.liu@uq.edu.au

Abbreviations: PPR: putative promoter region; TSS: transcription start site; Chr: chromosome (Chr22=chromosome 22); Ig: immunoglobulin; SW: starting window; O/E: observed/expected value; C, V, J, D, K and L: constant, variable, joining, diversity, kappa and lambda Ig genes; IGD: inter-gene distance

Introduction

The characterization of promoters and their regulatory elements is one of the major challenges in bioinformatics and functional genomics (Fickett and Hatzigeorgiou, 1997). Different approaches have been developed to detect conserved motifs in different genes (Fickett and Hatzigeorgiou, 1997; Fickett and Wasserman, 2000). Although the in silico approaches seem promising, unambiguous identification of regulatory elements is not straightforward (Fickett and Hatzigeorgiou, 1997). One problem is that the in silico approaches limit the putative promoter regions (PPRs) to an arbitrary number of base pairs upstream of the gene transcription start site (TSS). Ideally, this number should be chosen based on a functionally defined PPR because the length of the PPRs may differ considerably. Gene expression may be influenced not only by the regulatory sequences existed in the upstream and downstream flanking regions of a gene, possibly it is also influenced by the factors located in a remote distance (Shimada et al., 1989; Clegg et al., 1996). For example, the imprinting centre within human 15q11-q13 functions to co-regulate imprinted genes over a 2-Mb domain (Saitoh et al., 1996). X inactivation, as a regulatory mechanism related to gene expression, might influence a domain as far as >150 kb (Heard, 2004). Thus, to study the regulatory sequences of genes, a comprehensive analysis covering different numbers of base pairs upstream of the TSS of genes may be informative.

CpG islands are about 200-bp stretches of DNA that have a significantly higher concentration of CpG dinucleotides than the bulk of the genome (Davuluri et al., 2001; Ohlsson and Kanduri, 2002). CpG islands located in the gene PPRs play important roles in the reorganization of chromatin during mammalian spermiogenesis (Kundu and Rao, 1999) and in gene silencing during processes such as X-chromosome inactivation, imprinting, and silencing of intragenomic parasites (Takai and Jones, 2002). CpG islands are identified at the

5' end of approximately 60% of human genes and so are important genomic landmarks (Cross et al., 2000). Study of occurrence and characteristics of the CpG islands has gained great interests. Whole genome CpG island libraries have been prepared for human (Cross et al., 1994), chicken (McQueen et al., 1996), mouse (Cross et al., 1997) and pig (McQueen et al., 1997). These libraries provide a normalized set of sequences for the 5' end of CpG island-associated genes. Studies using these libraries have revealed that, in each species, CpG islands are not randomly distributed but are concentrated in particular regions (Cross et al., 2000). However, the mechanism of the CpG island in the regulation of gene expression remains unclear (Antequera, 2003). Little is known about whether and how density and organization of the CpG islands in gene promoter regions are gene-specific in human genome.

In vertebrates, antibody responses are one of the two classes of the immune responses to protect them from infection by microorganisms and parasites. The antibody responses involve the production of antibodies, which are proteins called immunoglobulins. An immunoglobulin (Ig) protein consists of two light chains (called as kappa and lambda chains, respectively) and two heavy chains. Interestingly, the genes (Ig genes) that express the three Ig chains are located on three chromosomes of both human and mouse genome. In human, they are located on Chr2, Chr14 and Chr22, while in mouse on Chr6, Chr12 and Chr16. Most of the Ig genes on human Chr2 and mouse Chr6 express the kappa chain and on human Chr22 and mouse Chr16 produce the lambda chain, while majority of the Ig genes from the human Chr14 and mouse Chr12 encode the Ig heavy chain. Recently, taking the in silico approach, we observed that CpG islands occur in the PPRs of the Ig genes on human Chr22 with very low frequency (Liu, 2005). Thus, we wonder if the Ig genes from human Chr2 and Chr14 also have a low frequency of the CpG islands occurred in their PPRs and whether the occurrence of the CpG islands in the PPRs is associated with the gene products (e.g., heavy,

kappa and lambda chains of the Ig). As little is known about whether and how CpG islands occur in the PPRs of the human Ig genes, we carried out a comprehensive analysis for the CpG islands in the PPRs of these genes. Here we describe the density, distribution pattern and organization of the CpG islands in the PPRs of the human and mouse Ig genes. Our results reveal that the characteristics of CpG islands in the PPRs of the Ig genes is highly different from that of the non-Ig genes on Chr22 in human genome.

Materials and Method

Both human and mouse gene-including DNA sequences were downloaded from the website of National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/mapview/) on and before 23rd, December 2004.

Sequence length: Each gene-including DNA sequence downloaded includes 5000bp upstream from the transcription start site (TSS), the full length of the gene (extron + intron) and the 1000bp downstream from the 3' end of the gene (Liu, 2005). If the inter-gene distance (the distance between the TSS of the gene being downloaded and the 3' end of the upstream gene) is less than 5000 bp, then the downloaded DNA sequence includes the actual sequence between the TSS of the gene and the 3' end of the upstream gene, the full length of the gene and 1000bp downstream from the 3' end of the gene. In each gene-including DNA sequence, the sequence upstream of the TSS is termed as a putative promoter region (PPR) and the sequence downstream of the 3' end of the gene is the 3' flanking region (Doyle and Han, 2001; Kemppainen et al., 2003).

Determination of the TSS: To determine the accurate position of the TSSs, each downloaded DNA sequence was carefully compared with the alignment of the same gene from Evidence

Viewer (EV) and Sequence Viewer (SV) of NCBI website. The downstream section of the PR (10-25 bases immediately upstream from the TSS) was aligned back with the corresponding section of the same gene illustrated on the EV and SV. Complete alignment confirmed correctness of the downloaded gene-including DNA sequence.

Criterion for identification of CpG island: The CpG island is a DNA sequence equal or longer than the starting window (SW), in which the observed/expected (O/E) value of CpGs should be equal or higher than 0.6 and G%+C%>50 (Gardiner-Garden and Frommer, 1987). Method for CpG island identification: A SW of 200 bp was applied from the 5' end of the DNA sequence, to compute the G%+C% and O/E value of the CpGs within the SW ((Liu, 2005), based on (Takai and Jones, 2002) with modification). Shifting the SW 1 bp towards downstream of the sequence after evaluation until the sequence within the SW met the criterion of the CpG island. Then elongating the window 1 bp on the front end (3' side) each time until the window no longer met the criterion. By taking 1 bp off from the 3' side of the window, the sequence within the window was considered as a CpG island. To exclude the mathematical CpG island (Liu, 2005), a condition was applied that at least 7 CpGs/200 bp should exist in the SW before a CpG island was reported. Two individual CpG islands were connected if they were separated by less than 50 bp. The CpG island information such as the total number of CpGs in each CpG island, the density of the CpG island (number of CpG/100bp), the geographic position of each CpG within the CpG island, the starting and ending coordinates of the CpG island along the gene sequence were collected for statistical analysis using Two-way ANOVA model.

Results

General description of the Ig genes:

Based on the data on 23rd, December 2004, 333 genes located on the three human chromosomes have been identified to express immunoglobulin (Ig) proteins, with 76 Ig genes on Chromosome 2 (Chr2), 170 on Chromosome 14 (Chr14) and 87 on Chromosome 22 (Chr22), respectively (Table 1 and 2). All the 76 Ig genes on Chr2 express Ig Kappa light chain, but only 1 gene (C Ig gene) encodes constant region and 5 short genes (J Ig gene, 38-39 bp) joining fragments between the variable and constant regions of the light chain, which are only located on the minus DNA strand. The other 70 Ig genes (V Ig gene), with 36 located on the minus strand and 34 on the plus strand, express Ig variable regions. The spatial distributions of the 76 Ig genes along the plus and minus strands can be simply described by two sketches (Sketches 1 and 2 in Table 1). The 170 Ig genes identified on Chr14 that express heavy Ig chain are all located on the minus strand (Table 1 and 2). Among them, there are 9 genes (J Ig gene) encoding the joining regions, 27 (D Ig gene) diversity regions, and 123 genes (V Ig gene) variable regions of the Ig heavy chain. The other 11 genes (C Ig gene) encode the constant regions of the heavy chain of alpha, gamma, delta, epsilon and mu Ig respectively. Their spatial distribution on Chr14 can be described by a sketch (Sketch 3 in Table 1). The 87 Ig genes found on Chr22 that express Ig lambda light chain are located only on the plus strand (Table 1 and 2), with 7 genes (C Ig gene) encoding the constant regions, 7 (J Ig gene) the joining regions and 73 (V Ig gene) variable regions. Their spatial distribution can be described by another sketch (Sketch 4 in Table 1). Here, an interesting phenomenon noted is the distribution order of the Ig genes along the DNA strand, both Chr2 and Chr14 exhibit to have the C Ig genes arranged at the 3' end of the minus strand, followed by the J and D Ig genes, with the V Ig genes distributed at the 5' end of the strand (Sketches 2 and 3, Table 1). Although the C and J Ig genes on Chr22 are arranged at the 3' end of the strand,

they are alternately distributed (Sketch 4, Table 1). Thus, the data may suggest that chromosomal location of the human Ig genes may determine the specific arrangement along the DNA strand.

Table 2 summarizes the length information of the Ig genes on the three chromosomes. On an average, the 27 D Ig genes on Chr14 that express the diversity region of the heavy chains have the shortest length (M=23bp). The V Ig genes encoding the variable regions have the average lengths of 590bp from Chr2, 376bp from Chr14 and 296bp from Chr22, which are significantly longer than the J Ig genes encoding the joining regions (38bp for Chr2 and Chr22; 55bp for Chr14). The lengths of the only 1 C Ig gene on Chr 2 and 7 C Ig genes on Chr22 that encode constant regions of the light chains are 323bp and 313bp, respectively, which are significantly shorter than those of 11 C Ig genes expressing constant regions of the heavy chains on Chr14 (2552bp, Table 2).

Inter-gene distance (IGD) represents the distance between the TSS of the gene and the 3' end of the upstream gene, which is another important physical character of a gene. The average IGD of V and J Ig genes is significantly longer on Chr22 than that on Chr2 and on Chr14 ($P<0.01$, Table 2). But, the averaged IGD of the 11 C Ig genes on Chr14 is significantly longer than that of 1 C Ig gene on Chr2 and 7 C Ig genes on Chr22 ($P<0.05$). Although the average IGD of the 5 J Ig genes on Chr2 is significantly shorter than that of the 7 J Ig genes on Chr22, but significantly longer than that of the 9 J Ig genes on Chr14 ($P<0.05$, Table2). The average IGD of the 73 V Ig genes on Chr22 is non-significantly longer than that of 123 V Ig genes on Chr14 and 70 V Ig genes on Chr22 ($P=0.1$, Table 2).

Frequency of the CpG islands in the PPRs of Ig genes on the three chromosomes:

Recently, we observed the frequency of the CpG island in the PPR of the Ig genes on Chr22 was apparently lower than that in the PPRs of the other genes (Liu, 2005). To investigate if

the sparse frequency of the CpG island in PPRs is a common phenomenon in all human Ig genes, we examined the occurrence of the CpG islands in the PPRs of all the Ig genes identified on the three chromosomes. As shown in Table 3, no CpG island was found in the PPRs of all the joining Ig genes irrespective of their chromosomal location. The only one constant Ig gene on Chr2 shows to have a CpG island (100%, 1/1), while 11 constant Ig genes from Chr14 have 6 CpG islands identified (55%, 6/11), but the 7 constant Ig genes from Chr22 show null CpG islands (0%, 0/7), indicating that the occurrence of the CpG islands in the PPRs of constant Ig genes depends on their chromosomal location. In terms of the frequency of the CpG islands in the PPRs of variable Ig genes, the frequency of the CpG island on Chr14 was slightly higher than those on the Chr2 and Chr22. The results suggest that the frequency of the CpG islands in the PPRs of variable Ig genes is also associated with chromosomal location. Comparing the frequency of the CpG islands occurred in the PPR of the Ig genes with that in the 439 Non-Ig genes (177% [779/439]) on Chr22, the CpG island in the PPR of Ig genes irrespective of chromosome localization is apparently sparse ($p=0.0008$). Fig. 1 and Fig. 2 show the occurrence of the CpG islands in the PPRs of 50 consecutively located non-Ig genes on Chr22 and 71 variable Ig genes, respectively. The frequency of the CpG island in the PPR of the Ig genes (Fig. 2) is significantly lower than that in the non-Ig genes (Fig. 1, $p<0.01$).

On Chr22, there are 79 genes, with inconclusive functions, named as LOC genes. Most of them are either hypothetical genes such as gene LOC128939 or pseudogenes. But some LOC genes can encode proteins that are structurally similar to a protein that has been identified in the cells. For example, gene LOC391284 produces the protein similar to neurobeachin. Here, we compared the localizations of the CpG islands identified in the PPRs of the Ig genes with those in the PPRs of the LOC genes and other genes (non-Ig and non-LOC genes encoding identified proteins) in Chr22. Apparently, the LOC genes have more

CpG islands in their PPRs overlapped with gene regions, with a frequency of 20% (10/50) that is significantly higher than the Ig genes (6%, 4/67), but significantly lower than the other genes(48%, 129/269) (Fig. 1, 3 and 4). Interestingly, all four Ig genes showing overlapped CpG islands encode the constant part of Ig heavy chain (Fig. 3), suggesting that the overlap of the CpG islands with TSS in the PPRs of Ig genes is related to their protein products.

Distribution and density of CpG islands in the PPRs of Ig and non-Ig genes:

We examined the distribution of the CpG islands in six length categories. The six length categories used were 200-250bp, 251-300bp, 301-400bp, 401-600bp, 601-1000bp and >1000bp. As a result, the Ig genes (55.9%) located on the three chromosomes and the other genes from Chr22 (66.2%) examined have their CpG islands dominantly distributed in the 200-250bp category (Fig. 5A). Meantime, 14.5% of Ig genes and 13.5% of other genes show to have the CpG islands fallen into the category of 401-600bp (Fig. 5A). Two-way ANOVA analysis indicates that the length distribution between two groups is non-significantly different ($P>0.2$).

We analyzed then the density of CpG island (CpGs/100bp) between Ig genes and the other genes on Chr22 using six CpG density categories, which were 3.5-4 CpG/100bp, 4.1-6 CpG/100bp, 6.1-8 CpG/100bp, 8.1-10 CpG/100bp, 10.1-15 CpG/100bp and >15 CpG/100bp. All the Ig genes on the three chromosomes have the density of CpG island less than 8 CpG/100bp, while the other genes on Chr22 have a wider range of the density of CpG island that occurred in all 5 categories, with 10.5% having the density between 8.1 to 15 CpG/100bp (Fig. 5B). Statistically, the CpG island density between the two groups of genes is significantly different ($p=0.00001$). Especially, 47% of the Ig genes had the CpG density in the category of 3.5-4 CpG/100bp, which was significantly higher than the other genes (34%)

(P<0.05). In contrast, only 4.2 % of the Ig genes had the CpG density in the category of 6.1-8 CpG/100bp, significantly lower than the other genes (11%) (P<0.01).

Clustered distribution pattern of CpGs within individual identified CpG islands in Ig and non-Ig genes:

We investigated next the clustered distribution pattern of the CpGs within individual identified CpG islands in the PPRs. Five clustered distribution patterns were adapted for this investigation. Pattern A (Pat A) is that the CpGs are randomly distributed within the island regions; Pattern B (Pat B) and C (Pat C) are those where the CpGs clustered on either 5' or 3' half of the island are 1.5 times denser than that on the other half, respectively; Pattern D (Pat D) is where the CpGs are mainly concentrated on two ends of the island, leaving the middle region of the island (>1/3 of the island length) blank; Pattern E (Pat E) is where the CpGs are mainly concentrated in the middle region of the island (Fig. 6). As shown in Fig. 5C, although Pat A is a dominant pattern in the two groups of genes, the non-Ig genes on Chr22 showed to have a significantly higher frequency (58.8%) than the Ig genes (37.8%) on the three chromosomes (P<0.01). Significant differences between the two groups of genes were also observed in Pat B and Pat D, with 21.6% of Ig genes in Pat B significantly higher than 8.2% of non-Ig genes (P<0.01), and 10.8% of Ig genes in Pat D significantly higher than 3.8% of other genes (P<0.01). In addition, the average lengths of the CpG island are 384 ± 570 bp identified in the Ig genes across the three chromosomes, which are significantly shorter than those (492 ± 592 bp) in the non-Ig genes on Chr22 (p<0.01), indicating that the CpG island length is also different between the two gene groups. But, the clustered distribution patterns were not related to the lengths of the CpG islands (P>0.1).

Occurrence of neighbouring CpGs:

Considering that the distance of 12 or 23 bp corresponds to approximately 1 or 2 helical turn(s) in the DNA sequence from a 3D point of view (Kim and Oettinger, 1998; Jones and Gellert, 2002), we examined how often the neighboring CpGs are aligned on the same side of the DNA 3D helix, and whether the frequency of occurrence of the neighboring CpG distance of 12 bp or 23 bp is higher than other distances. As a result, the occurrence of neighboring CpGs was not different between distances of 12 bp and 23 bp and other distances, suggesting that distribution of neighboring CpGs may not be gene specific.

CpG island distribution between the ATG^+ and ATG^- Ig genes:

We observed a large proportion of the Ig genes do not have the initiation codon ATG at their 5' end (ATG^- genes). It is also interesting to note that all the Ig genes with the ATG at their 5' end (ATG^+) are variable Ig genes (Table 4). Accordingly, we examined if the occurrence and distribution of the CpG islands are associated with the status of $ATG^{(+/-)}$. On Chr2, 55 out of 76 Ig genes (72%) encoding the kappa variable regions are ATG^+, which showed to have 11 CpG islands (20%) identified in their PPRs. 21 ATG^- Ig genes including 15 V, 5 J and 1 C Ig genes had 4 CpG islands (19%). Among the 170 Ig genes on Chr14, 79 ATG^+ Ig genes expressing variable regions were identified to have 21 CpG islands (26.5%, 21/79) in their PPRs. In the 91 ATG^- Ig genes including 11 C, 9 J, 27 D and 44 V Ig genes, 23 CpG islands (25%, 23/91) were identified. On Chr22, 87 Ig genes that are all ATG^- have 12 CpG islands in their PPRs (14%, 12/87). Thus, the results indicate that there is no relationship between the occurrence of the CpG islands in the PPRs and the existence of the initiation codon ATG in the Ig genes ($P>0.5$).

CpG island in mouse Ig genes:

Interestingly, 189 Ig genes identified in mouse genome are also located on three chromosomes, with 81 kappa genes on Chr6, 80 heavy genes on Chr12 and 28 lambda genes on Chr16, while the three chromosomes in human genome are Chr2, Chr14 and Chr22. However, only 25 mouse Ig genes (from 3 chromosomes) have their full length (PPR+gene+3' flanking region) of DNA sequences obtainable from the NCBI website at the time of composing this paper, with 5 CpG islands found in their PPRs (20%). No CpG islands are overlapped with the TSS (Fig. 7). The average lengths of the CpG islands in the mouse Ig genes are 251 ± 107 bp, significantly shorter than those in human Ig genes (384 ± 570 pb) ($p<0.001$), suggesting that determination of a CpG island length may be species-specific.

Discussions

The structure and function of immunoglobulin (Ig) genes have been studied for many years (Hozumi and Tonegawa, 1976; Gilmore-Hebert and Wall, 1978; Seidman et al., 1978; Max et al., 1979). Here, we examine first the physical structure of 333 human Ig genes located on Chr2, Chr14 and Chr22. Our data demonstrate that the physical characters of the human Ig genes including gene length, spatial distribution along the chromosomes and inter-gene distance differ greatly among the three chromosomes. Consequently, the Ig protein products expressed from the Ig genes are dependent on their chromosomal location, with that the Ig genes located on Chr2 produce Kappa light chain, on Chr14 heavy chain and on Chr22 lambda light chain. We examine then the characteristics of CpG islands in the putative promoter regions (PPRs) of human and mouse Ig genes. Our data have shown that 1). both human and mouse Ig genes are CpG islands poor in their PPRs and 2). CpG islands are not

randomly distributed but are concentrated in particular parts of the PPRs. Therefore, the density and organization of the CpGs in the PPR of a gene may reflect the species and gene specification.

In terms of CpG island identification in a gene using a computational method, two criteria have initially been used, which include (1) a DNA sequence longer than 200 bp (starting window, SW) with G+C content ≥ 50% and (2) Observed/Expected CpG ratio(O/E) ≥ 0.6 (Gardiner-Garden and Frommer, 1987). Based on these criteria, approximately 40% of genes were expected to be associated with CpG islands in animals (Gardiner-Garden and Frommer, 1987; Antequera and Bird, 1999). Using the two criteria with some modifications, Takai and Jones (Takai and Jones, 2002) analyzed the CpG islands on human Chr 21 and 22 although Marino-Ramirez et al. (Marino-Ramirez et al., 2004) argued that the length of DNA sequence of a CpG island should be at least 500 bp. Takai and Jones (Takai and Jones, 2002) observed that the distribution of the length of CpG islands in the 5' PPR showed a peak at the length range of 200-399bp, and a second smaller peak at the range of 1400-1599bp. We have used the similar criteria to examine the CpG islands for all the genes on human chromosome 22 (Liu, 2005) and all the Ig genes on three human chromosomes in the present study. We observed that the distribution of the CpG islands shows the peak at the length range of 200-250bp for both non-Ig genes on Chr22 and Ig genes on the three chromosomes. Thus, the present study, together with our recent results (Liu, 2005), indicates that a SW of 200 bp in length is preferable, supporting that the 200 bp SW is sensitive enough for determining the CpG islands in the PPRs (Takai and Jones, 2002). But, the length ranges defined in each of the six categories in the present study is narrower than those reported by Takai and Jones (Takai and Jones, 2002) because the length of the CpG islands in the PPRs of most Ig genes is relatively shorter than that of the non-Ig genes. Also, there is only one category for those

>1000bp in the present study, with a proportion of the CpG island for the category of >1000 much lower than those reported by Takai and Jones (Takai and Jones, 2002). This disparity is probably due to that we searched the CpG islands only from the PPRs of the genes while Takai and Jones analyzed the CpG islands along the whole gene sequences.

A question arising from the present study is why total number of CpG islands in the Ig genes (21.3%) is only about 12% of that in the non-Ig genes (177%, Table 3) on human Chr22. An early study has indicated that content of CpG islands in human genome is chromosome-dependent, with that Chr 18 is CpG island poor and Chr 22 CpG island rich (Cross et al., 2000). Recently, Antequera (Antequera, 2003) found that content of CpG islands in human genome is gene-dependent. He examined the occurrence of the CpG islands in PPRs of the genes that are transcribed by RNA polymerase II and found that the genes examined could be divided into two groups (Antequera, 2003). The first group of genes shows their expression restricted to a limited number of cell types to have a genome average frequency of the CpG islands in their PPRs (Antequera, 2003). In contrast, the second group of genes exhibits to have CpG density approximately 10 times higher (Antequera, 2003), which include all the housekeeping genes expressed in all cell types (Larsen et al., 1992; Antequera and Bird, 1993). In the present study, the results of rare CpG islands in their promoter regions of the Ig genes, associated with their cell-specific expression, support the observation by Antequera (Antequera, 2003). On the basis of DNA base composition analysis of the promoter regions of both Ig and non-Ig genes, we observed that the low density of CpG islands is associated with low percentage of CG contents in promoter regions of the Ig genes. Our unpublished data indicate that the average CG contents are only 44.95% in the PPRs of all Ig genes across the three chromosomes, which are apparently lower than 50.4% of CG contents in the PPRs of the genes on Chr 22. Thus, the remarkable deficiency of the CG dinucleotide may result in

loss of the CpG islands in promoter regions of the human Ig genes. Therefore, uneven occurrence of the CpG islands in promoter regions between Ig and non-Ig genes in human genome is observed in the present study. Meantime, loss of CpG islands may not be restricted to the human Ig genes, as similar situation has been observed for the mouse Ig genes. Based on the idea that CpG islands arose once at the dawn of vertebrate evolution, our data support the hypothesis for the sequence of CpG island evolution (Larsen et al., 1992; Antequera and Bird, 1993).

The clear results of the present work show that the average length of the CpG islands in Ig gene is shorter than that in non-Ig genes. The shorter length of the CpG islands in Ig genes is probably due to their lower CpG density because the computer program stops extending the searching window earlier in a region with lower CpG density than in a region with higher CpG density. But we can not exclude the possibility that short length of the CpG islands in a PPR is gene specific, which might be associated with gene expression and function.

In the present study, five clustered distribution patterns were adapted for examining the distribution of CpGs within an identified CpG island along the PPRs. Both Ig and non-Ig genes were grouped into the five patterns according to the position of the CpGs within the identified CpG islands. It is also clear that clustered distribution pattern of the CpGs varies considerably between non-Ig and Ig genes. High percentage of the genes belong only to one or two patterns. Our data have indicated that 58.8% of non-Ig genes exhibit random distribution of CpGs within the identified CpG islands (Pat A), which is significantly higher that the Ig genes (37.8%). In contrast, significantly higher percentage of the Ig genes than non-Ig genes have the clustered distribution of CpGs within the identified CpG islands belonging to Pat B and Pat D ($P<0.01$). Thus, clustered distribution of CpGs in the identified

CpG islands is gene-dependent, which may suggest that the clustered distribution of CpGs in PPRs is a useful landmark for distinguishing Ig genes from non-Ig genes in human genome.

CpG islands often overlap transcription units so they can be used to isolate full-length cDNAs for associated genes, either by probing cDNA libraries or by searching databases (Cross et al., 1999). Here, we observed that among the 67 Ig genes identified to have CpG islands in their PPRs, only 4 constant genes (6%) show the CpG islands overlapped with TSS, whereas significantly higher frequency of overlapped CpG islands was observed in non-Ig genes on Chr22 (48%). Cuadrado et al. (Cuadrado et al., 2001) observed that the overlap between CpG island and TSS is associated with two features in the identified CpG islands: methylation state and CpG-richness, which could have been generated by different mechanisms. Previous studies have suggested that the short unmethylated and CpG depleted regions at the 5' boundary of some islands do not fulfil the definition of CpG islands in terms of G+C content and CpG frequency (Bird, 1986; Gardiner-Garden and Frommer, 1987). As the present analysis can not determine methyl-CpGs within the identified CpG islands, we do not know whether the CpG islands in Ig genes is methylated or not and whether methylation of the CpG islands is different between Ig and non-Ig genes. Because methylation of CpG islands strongly influences both structural organization and function of chromatin (Kundu and Rao, 1999), a detailed study to examine methylation state of the CpG islands for both Ig and non-Ig genes is required.

Previous studies have clearly indicated that the total number of CpG islands in the mouse genome is about 20% less that in humans (Antequera and Bird, 1993; Matsuo et al., 1998). In the present study, the observed frequency of the CpG islands in the mouse Ig genes is lower than that of the human orthologous Ig genes. Meantime, the length of the CpG island in the

mouse Ig genes is significantly shorter than that of the human orthologous Ig genes. Similar situation between human and mouse genes has been previously observed. For example, the CpG island of the mouse adenine phosphoribosyltransferase (Aprt) gene extends approximately 80–100 bp upstream from the transcription initiation site (Dush et al., 1988; Macleod et al., 1994). In contrast, the 5' boundary of the CpG island of the human orthologue APRT lies 600 bp upstream from the transcription initiation site with an apparently higher CpG density (Antequera, 2003). Thus, the data support the conclusion that some human genes apparently have higher CpG density and longer CpG islands in their promoter regions than mouse orthologous genes (Antequera, 2003). In addition, human and mouse CpG islands are often differentially positioned or organized relative to the TSS at orthologous genes that are likely to play similar functions in both organisms, especially in the case of housekeeping genes (Cuadrado et al., 2001). It has been assumed that some human promoters 'acquired' a CpG island or some mouse promoters 'lost' it since the time they diverged in evolution (Antequera, 2003). However, according to the statement by Takai and Jones (Takai and Jones, 2002), CpG suppression in the human genome is caused not only by CpG depletion through evolution but also by the high content of simple repetitive sequences and low rate of sequence utilization for genes.

References

Antequera, F., 2003. Structure, function and evolution of CpG island promoters. Cell Mol Life Sci 60, 1647-58.

Antequera, F. and Bird, A., 1993. Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A 90, 11995-9.

Antequera, F. and Bird, A., 1999. CpG islands as genomic footprints of promoters that are associated with replication origins. Curr Biol 9, R661-7.

Bird, A.P., 1986. CpG-rich islands and the function of DNA methylation. Nature 321, 209-13.

Clegg, C.H., Haugen, H.S. and Boring, L.F., 1996. Promoter sequences in the RI beta subunit gene of cAMP-dependent protein kinase required for transgene expression in mouse brain. J Biol Chem 271, 1638-1644.

Cross, S.H., Charlton, J.A., Nan, X. and Bird, A.P., 1994. Purification of CpG islands using a methylated DNA binding column. Nat Genet 6, 236-44.

Cross, S.H., Clark, V.H. and Bird, A.P., 1999. Isolation of CpG islands from large genomic clones. Nucleic Acids Res 27, 2099-107.
Cross, S.H., Clark, V.H., Simmen, M.W., Bickmore, W.A., Maroon, H., Langford, C.F., Carter, N.P. and Bird, A.P., 2000. CpG island libraries from human chromosomes 18 and 22: landmarks for novel genes. Mamm Genome 11, 373-83.
Cross, S.H., Lee, M., Clark, V.H., Craig, J.M., Bird, A.P. and Bickmore, W.A., 1997. The chromosomal distribution of CpG islands in the mouse: evidence for genome scrambling in the rodent lineage. Genomics 40, 454-61.
Cuadrado, M., Sacristan, M. and Antequera, F., 2001. Species-specific organization of CpG island promoters at mammalian homologous genes. EMBO Rep 2, 586-92.
Davuluri, R.V., Grosse, I. and Zhang, M.Q., 2001. Computational identification of promoters and first exons in the human genome. Nat Genet 29, 412-7.
Doyle, M.C. and Han, I.S., 2001. The roles of two TATA boxes and 3'-flanking region of soybean beta-tubulin gene (tubB1) in light-sensitive expression. Mol Cells 12, 197-203.
Dush, M.K., Briggs, M.R., Royce, M.E., Schaff, D.A., Khan, S.A., Tischfield, J.A. and Stambrook, P.J., 1988. Identification of DNA sequences required for mouse APRT gene expression. Nucleic Acids Res 16, 8509-24.
Fickett, J.W. and Hatzigeorgiou, A.G., 1997. Eukaryotic promoter recognition. Genome Res 7, 861-878.
Fickett, J.W. and Wasserman, W.W., 2000. Discovery and modeling of transcriptional regulatory regions. Curr Opin Biotechnol 11, 19-24.
Gardiner-Garden, M. and Frommer, M., 1987. CpG islands in vertebrate genomes. J Mol Biol 196, 261-282.
Gilmore-Hebert, M. and Wall, R., 1978. Immunoglobulin light chain mRNA is processed from large nuclear RNA. Proc Natl Acad Sci U S A 75, 342-5.
Heard, E., 2004. Recent advances in X-chromosome inactivation. Curr Opin Cell Biol 16, 247-255.
Hozumi, N. and Tonegawa, S., 1976. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. Proc Natl Acad Sci U S A 73, 3628-32.
Jones, J.M. and Gellert, M., 2002. Ordered assembly of the V(D)J synaptic complex ensures accurate recombination. Embo J 21, 4162-71.
Kemppainen, R.J., Cox, E., Behrend, E.N., Brogan, M.D. and Ammons, J.M., 2003. Identification of a glucocorticoid response element in the 3'-flanking region of the human Dexras1 gene. Biochim Biophys Acta 1627, 85-9.
Kim, D.R. and Oettinger, M.A., 1998. Functional analysis of coordinated cleavage in V(D)J recombination. Mol Cell Biol 18, 4679-88.
Kundu, T.K. and Rao, M.R., 1999. CpG islands in chromatin organization and gene expression. J Biochem (Tokyo) 125, 217-22.
Larsen, F., Gundersen, G., Lopez, R. and Prydz, H., 1992. CpG islands as gene markers in the human genome. Genomics 13, 1095-107.
Liu, G.B., Jiang, Y.F., Yan, H., Yan, J.Q., Pettigrew, J.D. Zhao, K.N., 2005. Physical characteristics of gene putative promoter regions in human chromosome 22. submitted.
Macleod, D., Charlton, J., Mullins, J. and Bird, A.P., 1994. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev 8, 2282-92.

Marino-Ramirez, L., Spouge, J.L., Kanga, G.C. and Landsman, D., 2004. Statistical analysis of over-represented words in human promoter sequences. Nucleic Acids Res 32, 949-958.
Matsuo, K., Silke, J., Georgiev, O., Marti, P., Giovannini, N. and Rungger, D., 1998. An embryonic demethylation mechanism involving binding of transcription factors to replicating DNA. Embo J 17, 1446-53.
Max, E.E., Seidman, J.G. and Leder, P., 1979. Sequences of five potential recombination sites encoded close to an immunoglobulin kappa constant region gene. Proc Natl Acad Sci U S A 76, 3450-4.
McQueen, H.A., Clark, V.H., Bird, A.P., Yerle, M. and Archibald, A.L., 1997. CpG islands of the pig. Genome Res 7, 924-31.
McQueen, H.A., Fantes, J., Cross, S.H., Clark, V.H., Archibald, A.L. and Bird, A.P., 1996. CpG islands of chicken are concentrated on microchromosomes. Nat Genet 12, 321-4.
Ohlsson, R. and Kanduri, C., 2002. New twists on the epigenetics of CpG islands. Genome Res 12, 525-6.
Saitoh, S., Buiting, K., Rogan, P.K., Buxton, J.L., Driscoll, D.J., Arnemann, J., Konig, R., Malcolm, S., Horsthemke, B. and Nicholls, R.D., 1996. Minimal definition of the imprinting center and fixation of chromosome 15q11-q13 epigenotype by imprinting mutations. Proc Natl Acad Sci U S A 93, 7811-7815.
Seidman, J.G., Leder, A., Nau, M., Norman, B. and Leder, P., 1978. Antibody diversity. Science 202, 11-7.
Shimada, T., Fujii, H. and Lin, H., 1989. A 165-base pair sequence between the dihydrofolate reductase gene and the divergently transcribed upstream gene is sufficient for bidirectional transcriptional activity. J Biol Chem 264, 20171-20174.
Takai, D. and Jones, P.A., 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A 99, 3740-3745.

Legends:

Figure 1. Distribution of the CpG island in the PPRs of 50 non-Ig genes on Chr22. The CpG islands in the PPRs of 27 genes (54%) are overlapped with the TSS (those located at the right hand side of the PPRs).

Figure 2. Distribution of the CpG islands in the PPRs of the 71 human Ig variable genes on Chr22. The CpG islands are apparently sparse in the PPRs of this group of genes.

Figure 3. Distribution of the CpG islands in the PPRs of 67 human Ig Genes where the CpG islands were found. The top 10 PRs are for the Ig constant genes on Chr14, among which 4

CpG islands (top-right) are overlapped with the TSS. None of the other Ig genes have the overlapped CpG islands.

Figure 4. Distribution of the CpG islands in the PPRs of 50 LOC genes on Chr22. Twenty percentage of the CpG islands (10/50) in the PPRs are overlapped with the TSS.

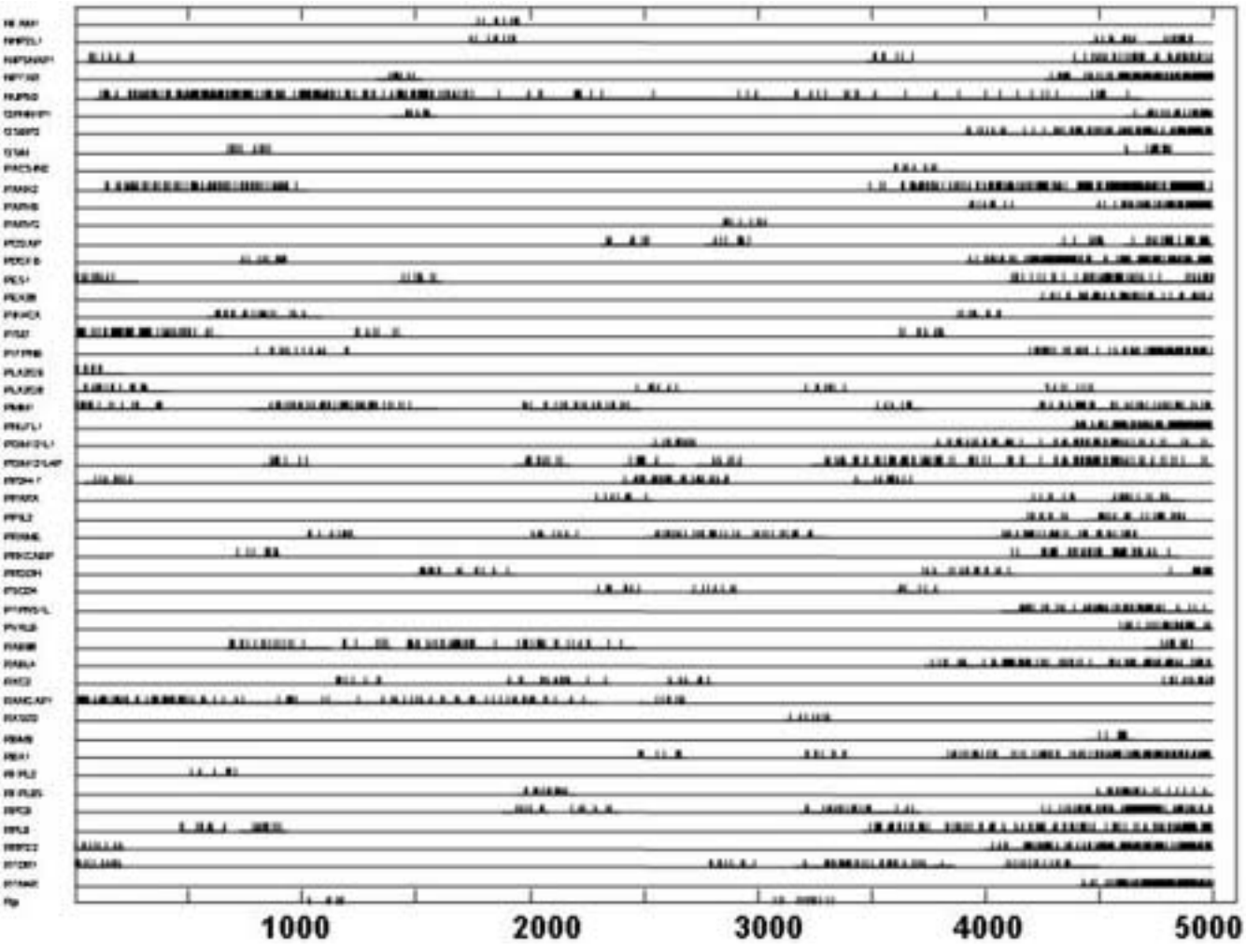
Figure 5. Distribution of the length (A), density (B) and pattern (C) of the CpG Islands for human Ig genes across 3 chromosomes and other (non-Ig and non-LOC) genes on Chr22. Statistical test: *, $P<0.05$; **, $P<0.01$.

Figure 6. Clustered distribution pattern of CpGs within a CpG island. **Pat A**: the CpGs are randomly distributed. **Pat B and C**: the CpGs in the 5' or 3' end of the island are 1.5 times denser than those in the other end respectively. **Pat D**: CpGs are dense in two ends of the island, with the middle sequence of >1/3 of the island length being blank. **Pat E**: CpGs are dense in the middle region of the island.

Figure 7. Distribution of the CpG island in the PPRs of 25 mouse Ig genes. The length of the CpG island in the PPRs of the mouse Ig genes is signficantly shorter than that of the human Ig genes ($P<0.01$).

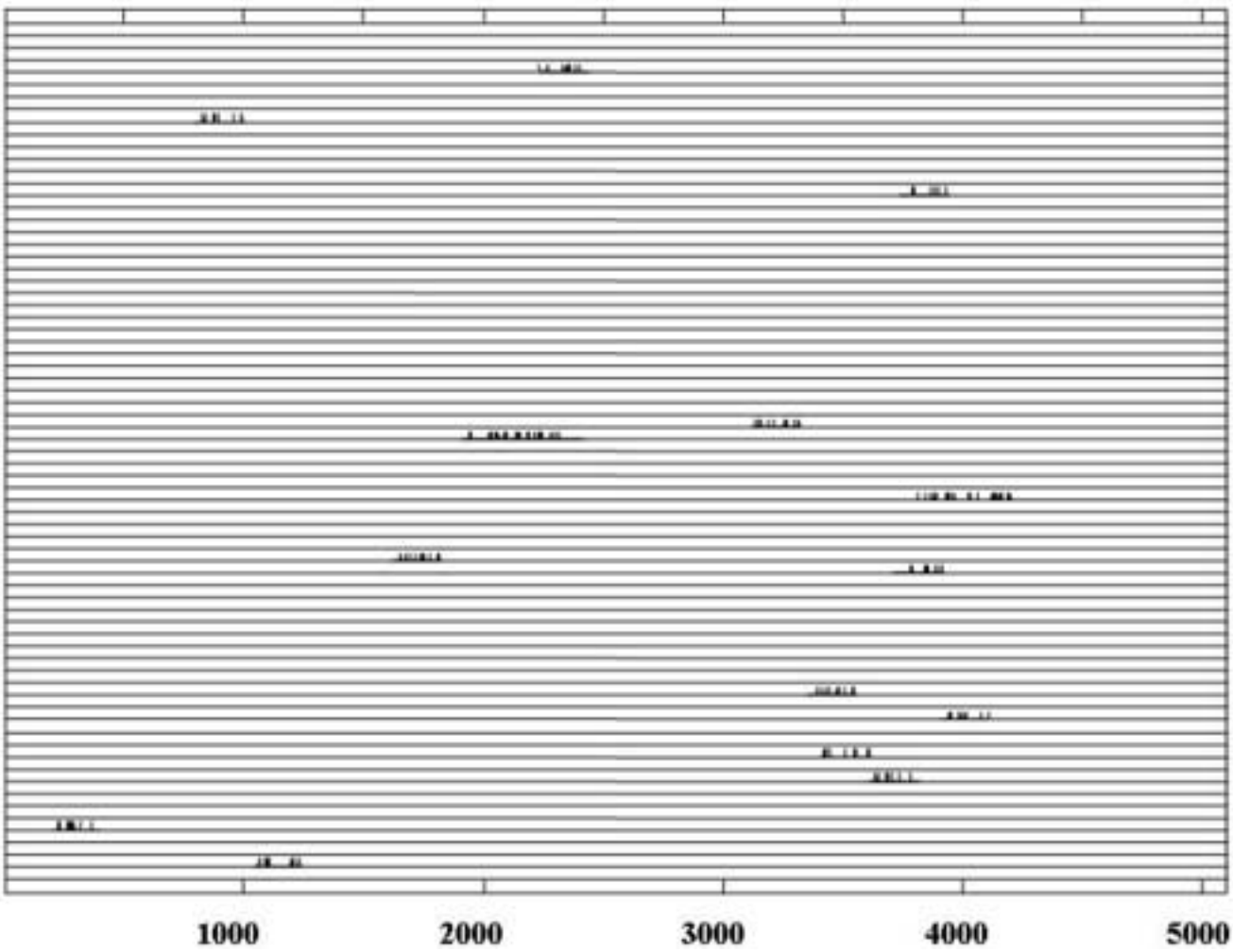
Figure(1)

Click here to download high resolution image



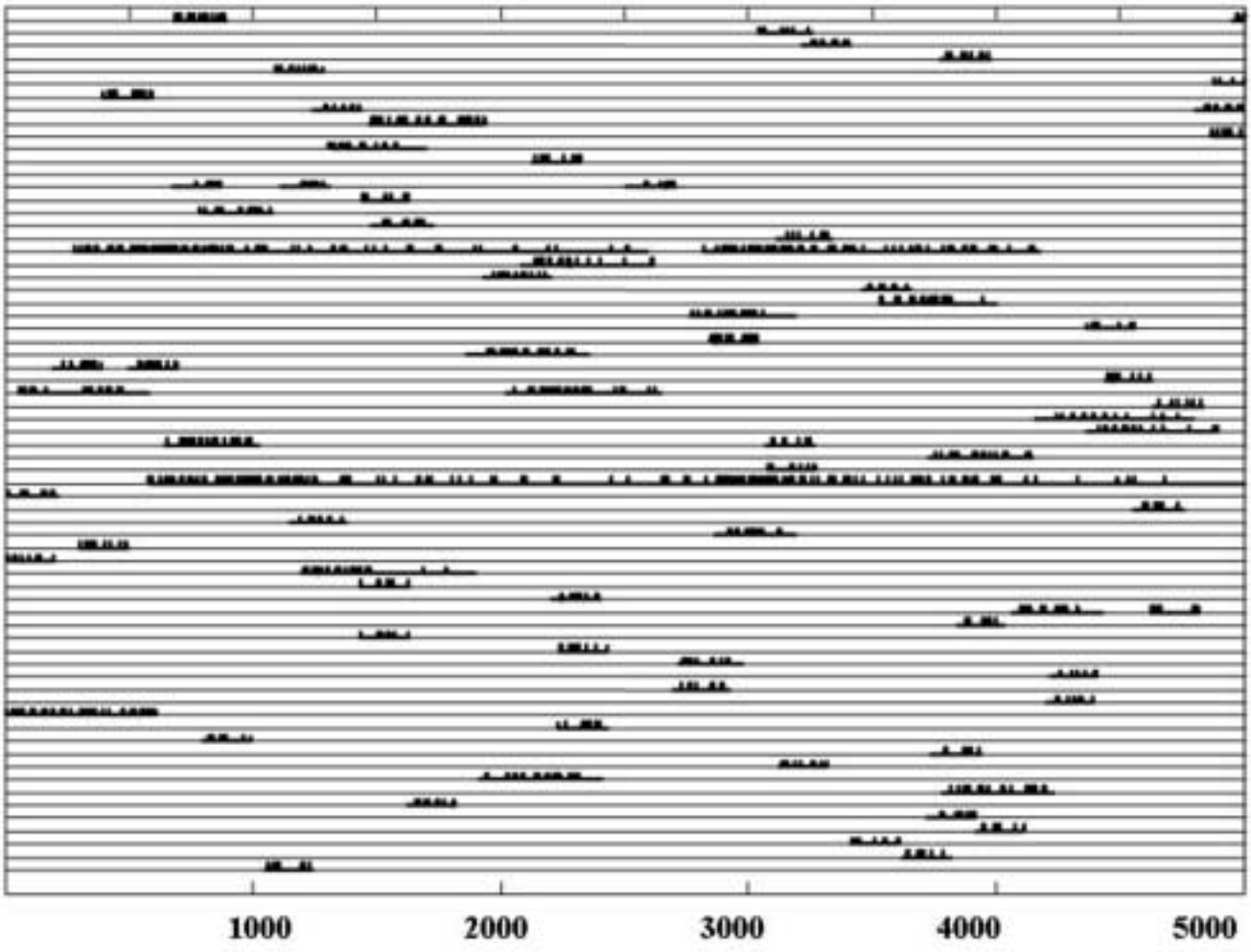
Figure(2)

Click here to download high resolution image



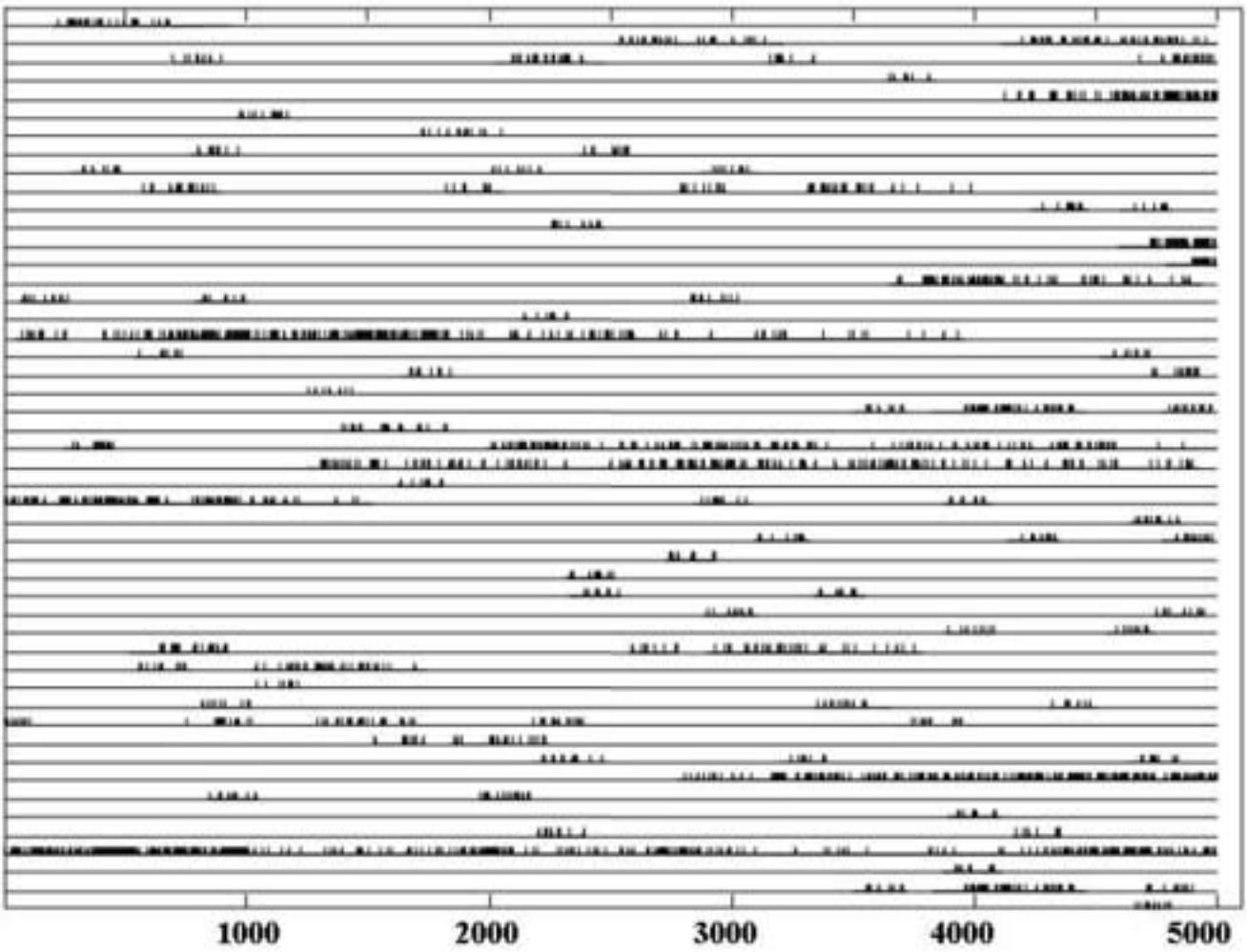
Figure(3)

Click here to download high resolution image

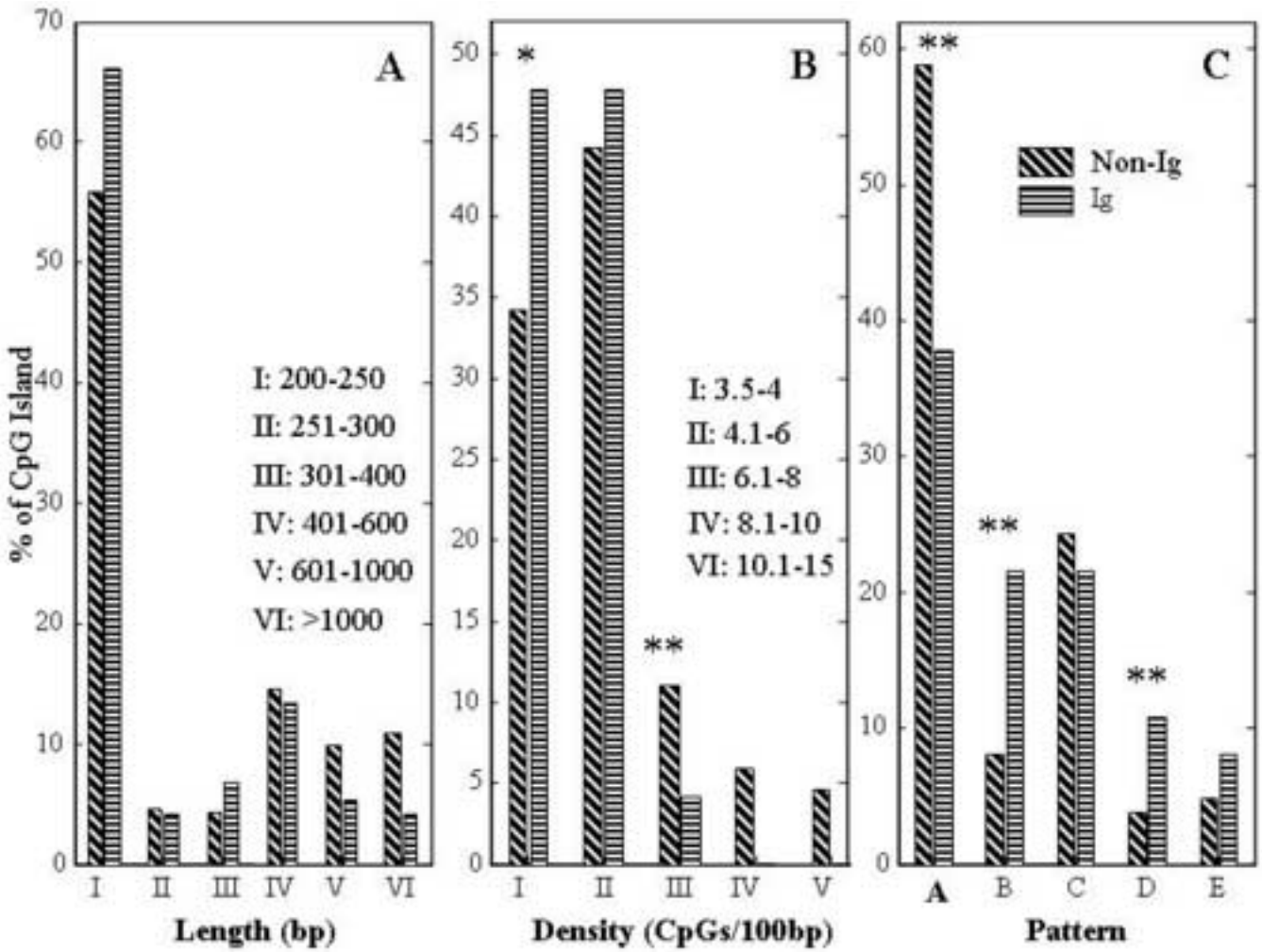


Figure(4)

Click here to download high resolution image

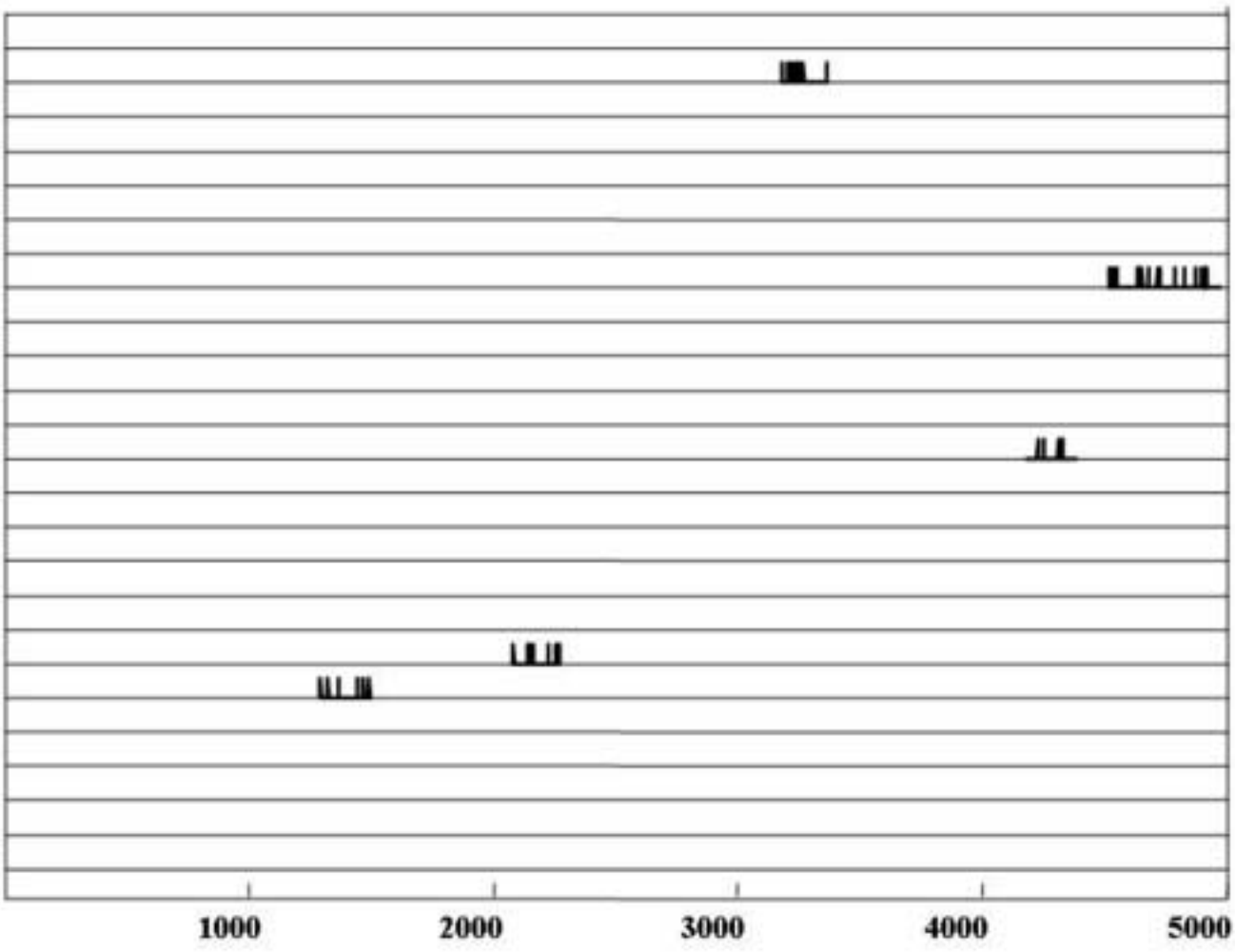


Figure(5)
Click here to download high resolution image



Figure(6)
Click here to download high resolution image

Figure(7)
Click here to download high resolution image



Table(1)

Table 1. Schematic representation of the Ig gene distributions on 3 chromosomes (numbers are the number of genes in each sub-group)

Sketch	Chr	Strand	Gene distribution (chain)
1	2	+	5’ --------V_{2}-----V_{32}---------- 3’ (kappa)
2	2	-	3’ ---$C_{1}J_{5}$----V_{36}---------- 5’ (kappa)
3	14	-	3’ ------$C_{11}J_{9}D_{27}V_{123}$-------- 5’ (heavy)
4	22	+	5’-----$V_{73}J_{1}C_{1}J_{1}C_{1}J_{1}C_{1}J_{1}C_{1}J_{1}C_{1}J_{1}C_{1}J_{1}C_{1}$-----3’ (lambda)

Table 2. General information of the confirmed genes encoding different chains of the immunoglobulin (based on data available 13th July 2004).

Chromosome	Group Name (strand)	No of Genes	Mean Size (Min-Max)	Inter-gene Distance Mean (Min-Max)
Chr2	kappa constant (-)	1	323	2850
	kappa joining (-)	5	38 (38-39)	290 (265-320)
	kappa variable (36-; 34+)	70	590 (211-1778)	60 genes: >5K 10 genes: 1684 (805-3284)
Chr14	heavy constant (-)	11	2522	10 genes >5K 1 gene = 1bp
	heavy joining (-)	9	55 (50-65)	239 (92-348)
	heavy diversity (-)	27	23 (11-37)	1460 (62-2666)
	heavy variable (-)	123	376 (77-501)	62 genes: >5K 61 genes: 2775 (559-4937)
Chr22	lambda constant (+)	7	313 (253-353)	1 gene: >5K 6 genes: 3081 (1539-3923)
	lambda joining (+)	7	38 (38-38)	1173 (220-1558)
	lambda variable (+)	73	296 (280-324)	42 genes: >5K 31 genes: 3079 (791-4652)

Legends:

Group Name (strand): the name of the gene group indicating their functions and which strand they are located (+ / -).

No of genes: the number of genes found in the corresponding gene group.

Mean Size (Min-Max): The mean size of the gene in the corresponding group. The numbers in parentheses indicate the minimal and maximal size of the genes in the group.

Inter-gene distance (Min-Max): The tail-head distance between the end of the upstream gene and the gene being analysed. The numbers in parentheses indicate the minimal and maximal tail-head distances in the group.

Table(3)

Table 3. The frequency of the CpG island on the promoter regions of different groups of Ig genes and the Non-Ig genes on Chr22.

Chr	Group Name	Gene No	Length of PR (bp)	CpG Isl	CpG Isl frequency %
2 (K)	Constant	1	2580	1	100
	Joining	5	290 (M)	0	0
	Variable	60	5000	15	25
	Variable	10	1700 (M)	0	0
14 (H)	Constant	11	5000	6	33
	Joining	9	239 (M)	0	0
	Diversity	27	1460 (M)	2	7
	Variable	62	5000	23	37
	Variable	61	2775 (M)	12	20
22 (L)	Constant	7	3081 (M)	0	0
	Joining	7	1173 (M)	0	0
	Variable	42	5000	5	12
	Variable	31	3079 (M)	7	23
Total Ig Gene		333	CpG Isl found	71	21.3
Non-Ig genes on Chr22		439	CpG Isl found	779	177

Chr: Chromosome name. The letters in parentheses indicate the Ig chain the group of Ig genes are encoding. K=keppa, H=heavy, L=lambda.
Group Name: the name of the gene groups implicating their functions.
Gene No.: The number of genes in each group.
Length of PR: The length of the promoter regions. For those that are shorter than 5000bp, only the mean length (M) is shown.
CpG Isl: The number of CpG island found from the corresponding gene group.
CpG Isl frequency %: the number of CpG island found on every 100 PRs in each group.

Table(4)

Table 4. Distribution of the ATG+ and ATG- Ig genes in the 3 chromosomes.

Chromosome	No of Ig Genes	ATG+ genes (%)	ATG- Genes (%)
Chr2	76	55 (72%) All Variable	21 (28%)
Chr14	170	79 (46.5%) All Variable	91 (53.5)
Chr22	87	nill	87 (100)